HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation
Spring 2015

Topic 4: GREG and calibration estimators
PART I: Direct GREG and calibration

Risto Lehtonen, University of Helsinki

Topic 4 Part I

- **GREG and calibration estimators**
  PART I: Direct GREG and calibration
  - Introductory remarks
  - Difference estimator
  - Population fit regression estimator
  - Direct GREG estimator for domain totals
  - Variance estimators
  - Example

Survey-driven data infrastructure

- **"Survey" countries**
  - Restricted availability of unit-level administrative registers for statistical purposes
  - Official statistics production mainly based on sample surveys
  - Data from different sources can most often be merged at an aggregate level only
  - Area-level models and area-level auxiliary data are often used for estimation for domains and small areas

Register-driven data infrastructure

- **"Register" countries**
  - Administrative and statistical registers are available for statistical purposes
  - Large share of official statistics production are based on statistical registers
  - Sample surveys complete the register data sources
  - Unique ID codes available in the various data sources
  - Micro-merging of data from different sources possible
  - Unit-level models and unit-level auxiliary data are often used for estimation for domains and small areas

Approaches chosen

- **Both survey-driven and register-driven data infrastructures are discussed**
  - Survey-driven: Flexible use of sample data and aggregated auxiliary data (e.g. GREG and calibration)
  - Register-driven: Combined use of sample survey data and unit-level auxiliary register data (extended GREG)
- **Additional points favouring register-driven**
  - Many countries in Europe and elsewhere are using, or are turning towards, register-driven infrastructure
  - Much research is on-going under this option

Components of estimation procedure under register-driven option

- **Sample survey data**
  - Access to unit-level sample survey data
  - Model specification and model fitting
- **Auxiliary data**
  - Access to auxiliary unit-level population data
- **SAE and domain estimation**
  - Merging of sample data and auxiliary data at unit level
  - Calculation of fitted values for all population elements
  - Calculation of domain estimators and accuracy measures

Population fit regression estimator - 1

Difference estimator of population total t of y (Särndal 1980)

Let us assume known values y_k^0 that are close to population values y_k, $k \in U$. We write the population total as

$$t = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0) \qquad (8)$$

Assume that sample values y_k, $k \in s$ only are available

Difference estimator: We estimate the second sum using HT:

$$\hat{t}_{DIFF} = \sum_{k \in U} y_k^0 + \sum_{k \in s} a_k (y_k - y_k^0), \text{ where } a_k = 1/\pi_k$$

In practice, no such y_k^0, $k \in U$, exist! Let us use modelling...

Population fit regression estimator - 2

Consider regression superpopulation model

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, \quad Var(\varepsilon_k) = \sigma_k^2 = \sigma^2 \text{ (constant)}$$

where $\mathbf{x}_k = (1, x_{1k}, ..., x_{Jk})'$ is the vector of auxiliary x-variables

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$ is the vector of regression coefficients

If we had access to population values $y_k \in U$ then a GLS (generalized least squares) estimator $\tilde{\mathbf{B}}_{GLS}$ of $\boldsymbol{\beta}$ is:

$$\tilde{\mathbf{B}}_{GLS} = \left(\sum_{k \in U} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in U} a_k \mathbf{x}_k y_k \right)$$

NOTE: $\tilde{\mathbf{B}}_{GLS}$ is determined at the population level

Population fit regression estimator - 3

Using $\tilde{\mathbf{B}}_{GLS}$ and $\mathbf{x}_k$, $k \in U$, we calculate fitted values $\tilde{y}_k = \mathbf{x}_k' \tilde{\mathbf{B}}_{GLS}$ for all $k \in U$. We define **population fit regression estimator :**

$$\hat{t}_{REG} = \sum_{k \in U} \tilde{y}_k + \sum_{k \in S} a_k (y_k - \tilde{y}_k), \quad \text{where} \quad a_k = 1/\pi_k$$

In practice we only have access to sample values $y_k \in s$

Thus, we estimate **B** by plugging in HT estimators for both sum components (weighted least squares estimator):

$$\hat{\mathbf{B}}_{WLS} = \left(\sum_{k \in S} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in S} a_k \mathbf{x}_k y_k \right) \text{ and } \hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}$$

Direct GREG estimator for domains - 1

GREG is a sample-based substitute for the population fit regression estimator

Direct GREG estimator of domain total $t_d = \sum_{k \in U_d} y_k$:

Assisting model:

$$y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k, \; Var(\varepsilon_k) = \sigma^2$$

Domain-specific parameter $\mathbf{B}_d$ is estimated by

$$\hat{\mathbf{B}}_d = \left(\sum_{k \in s_d} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s_d} a_k \mathbf{x}_k y_k$$

Direct GREG estimator for domains - 2

Fitted values $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_d$ and residuals $e_k = y_k - \hat{y}_k$

are incorporated into the **direct GREG estimator**

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k \quad (9)$$

First part: Synthetic (SYN) estimator

Second part: HT estimator of residual total $\sum_{k \in U_d} E_k$

(adjustment for design bias of SYN estimator)

NOTE: Adjustment term sometimes = zero

More detailed: See Lehtonen-Veijanen (2009) pp. 228-230
(separate pdf sheet on course website)

Traditional regression estimator

Rearranging the terms of GREG: traditional regression estimator

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{B}}_d, \qquad (10)$$

where $\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k = \left(N_d, \sum_{k \in U_d} \mathbf{x}_{1k}, \ldots, \sum_{k \in U_d} \mathbf{x}_{Jk}\right)'$

$$\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} a_k \mathbf{x}_k$$

Variance of $\hat{t}_{dGREG}$ can be approximated using sample residuals $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_d$:

$$\hat{V}_1\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) e_k e_l \qquad (11)$$

GREG using g-weights

GREG can be written as a weighted sum of observations incorporating so-called g-weights (Särndal et al. 1992):

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k, \qquad (12)$$

where $g_{dk} = I_{dk} + I_{dk} \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{M}}_d^{-1} \mathbf{x}_k$ and $\hat{\mathbf{M}}_d = \sum_{i \in s_d} a_i \mathbf{x}_i \mathbf{x}_i'$

$I_{dk} = I\{k \in U_d\}$ is the domain membership indicator

g-weights are used in variance estimator

$$\hat{V}_2 \left(\hat{t}_{dGREG} \right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (13)$$

Calibration property

GREG as calibration estimator

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k$$

Calibration for auxiliary x-variables

$$\hat{t}_{dGREG}(x_j) = \sum_{k \in s_d} a_k g_{dk} x_{jk} = t_{dx} = \sum_{k \in U_d} x_{jk} \text{ , } j=1,...,J$$

Applying g-weights for any x-variable reproduces the known population total of x-variable in domain *d*

EXAMPLE: Ratio-type SYN and GREG

Continuous y-variable and auxiliary x-variable

Domains of interest U_d, $d = 1,...,D$

Assisting model (linear fixed-effects model)

$$y_k = \beta_1 x_k + \varepsilon_k, \quad k \in U$$

NOTE: Intercept parameter $\beta_0 = 0$

Direct and indirect SYN estimators

Direct and indirect GREG estimators

Direct SYN estimator

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{B}_{1d} x_k = t_{dx} \hat{B}_{1d} = t_{dx} \times \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}$$

$$= \sum_{k \in U_d} x_k \times \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k x_k}$$

NOTE: This SYN estimator is direct - Why?

Indirect SYN estimator

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{B}_1 x_k = t_{dx} \hat{B}_1 = t_{dx} \times \frac{\hat{t}_{HT}}{\hat{t}_{xHT}}$$

$$= \sum_{k \in U_d} x_k \times \frac{\sum_{k \in s} a_k y_k}{\sum_{k \in s} a_k x_k}$$

NOTE: This SYN estimator is indirect - Why?

Bias of indirect SYN estimator

$$\mathrm{BIAS}(\hat{t}_{dSYN}) = E(\hat{t}_{dSYN}) - t_d \doteq -t_{dx}(B_{1d} - B_1), \text{ where}$$

$$B_{1d} = \sum_{k \in U_d} y_k \,/\, \sum_{k \in U_d} x_k$$

is the domain-specific slope parameter

$$B_1 = \sum_{k \in U} y_k \,/\, \sum_{k \in U} x_k$$

is the slope defined at the population level

Bias is small if domain slopes B_{1d} are close to the population slope B_1 - if not, bias can be large!

Direct GREG estimator

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \hat{t}_{dSYN} + \sum_{k \in s_d} a_k (y_k - \hat{B}_{1d} x_k)$$

$$= \hat{t}_{dHT} + \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}} (t_{dx} - \hat{t}_{dxHT}) = t_{dx} \times \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}$$

NOTE: This direct GREG estimator is the same as the direct SYN!

Indirect GREG estimator

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \hat{t}_{dSYN} + \sum\nolimits_{k \in s_d} a_k (y_k - \hat{B}_1 x_k)$$

$$= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{xHT}} (t_{dx} - \hat{t}_{dxHT})$$

NOTE: This GREG is indirect - Why?

EXAMPLE
Direct HT and direct GREG for planned domains

- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

- Section 3.5. Computational example with direct estimation under a planned domain structure

Sampling design

- Population: N = 431,000 households
- Household sampling: Stratified πPS (PPS-WOR)
- Size variable in PPS-WOR: Number of household members
- Strata: D = 12 NUTS4 regions (domains)
- Planned type domains
- Proportional allocation
  - Domain (stratum) sample sizes are assumed fixed
- Total sample size: n = 1000 households

Variables

- **Study variable** *y*
  - Disposable household income
- **Auxiliary x-variables (known for all HHs)**
  - EDUC: the number of household members who had higher education
  - EMP: the number of months in total the household members were employed during last year
  - Variables are derived from administrative registers
- NOTE: for this pedagogical exercise we also assume access to population values for study variable y
- This gives option to compare results with true values

HT estimator

HT estimator $\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k = \sum_{k \in s_d} y_k \, / \, \pi_K$

For variance estimation, we approximate the design by with-replacement type PPS (SAS Procedure SURVEYMEANS)

Variance estimator (4) for planned domains

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2$$

Direct GREG estimator

Alternative formulations:

Eq. (9)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

Eq. (10)

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{B}}_d$$

Eq. (12)

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k$$ (calibration estimator)

Practical variance estimator for direct GREG for planned domains

Approximate variance estimator of GREG:

$$\hat{V}_A\left(\hat{t}_{dGREG}\right) = \frac{1}{n_d(n_d-1)}\sum_{k\in s_d}\left(n_d a_k e_k - \hat{t}_{dHTe}\right)^2 \qquad (14)$$

where

n is the total sample size and $a_k = 1/\pi_k$ (design weights)

$e_k = y_k - \hat{y}_k$ are residuals in fitting the model

$y_k = \beta_{0d} + \beta_{1d}x_{1k} + \beta_{2d}x_{2k} + ... + \beta_{Jd}x_{Jk} + \varepsilon_k, \quad k \in U_d$

$\hat{t}_{dHTe} = \sum_{k\in s_d} a_k e_k$ is HT estimator of residual total in domain d

NOTE: Similarity of (14) with HT variance estimator (4) for planned domains
Both (4) and (14) are used in RDomest software

Assisting models in GREG

Direct GREG estimator with linear fixed-effects assisting model and domain-specific terms

$$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \varepsilon_k \quad \text{(column 2), or}$$

$$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \beta_{2d}\mathrm{EDUC}_k + \varepsilon_k \quad \text{(column 3)}$$

NOTE: Domain-specific intercepts and slopes

Quality measures of estimators

ARE Absolute relative error of an estimator in domain *d*

$$\mathrm{ARE}(\hat{t}_d) = |\hat{t}_d - t_d| / t_d \text{ , } d = 1,...,D$$

MARE in domain group:
The mean of absolute relative errors over domains in the group

MCV The mean coefficient of variation of the estimate over domain group

The coefficient of variation is calculated as $s.e(\hat{t}_d) / \hat{t}_d$

where s.e refers to the estimated standard error of an estimator

Table 3. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and direct calibration (GREG) estimators of totals for minor, medium-sized and major domains by using various amounts of auxiliary information for **planned domains**.

	HT		GREG			
	Auxiliary information					
	1 None		2 Domain sizes and domain totals of EMP		3 Domain sizes and domain totals of EMP and EDUC	
Domain sample size class	MARE %	MCV %	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	**11.9**	5.8	**7.7**	6.4	**6.8**
Medium $34 \le n_d \le 45$	7.6	**9.0**	3.7	**8.0**	3.6	**8.1**
Major $46 \le n_d \le 277$	12.5	**5.2**	4.3	**4.7**	5.2	**3.7**

Lessons learned – Planned domains

- **Estimation error**
  - Mean absolute error MARE figures are smaller for GREGs when compared with HT, in all three domain sample size groups
- **Estimation accuracy (variance)**
  - Mean coefficient of variation MCV figures tend to be smaller for both GREGs, when compared with HT
  - GREG with more use of auxiliary data tends to be more accurate than the GREG with less use of auxiliary data
- Incorporation of auxiliary data in the estimation procedure makes sense